Les sondages : délaissés par les statisticiens et malmenés par les politologues

Jeanne Fine(*) & Jean-Louis Piednoir(**)

Les périodes électorales sont l'occasion de s'interroger sur une technique qui tient à la fois de la sociologie et de la statistique, donc des mathématiques : les sondages. Mais il y a des sondages tout au long de l'année, pour les cotes de popularité, pour le lancement d'un nouveau produit de grande consommation. Pour une manifestation organisée par la société française de statistique (SFdS), Jeanne Fine a rédigé cet article de vulgarisation qui a été revu et complété par Jean Louis Piednoir qui la remercie vivement de l'avoir autorisé à adapter son texte.

1. Introduction, vocabulaire

Le point de départ de ce travail est la publication dans *Le Monde* du 26 avril 2002 de deux articles situés côte à côte : « *Sondages et regrets* » par Roland Cayrol, Directeur associé de l'Institut CSA, et « *Faute de contrôles* » par Michel Lejeune, Statisticien. Les deux auteurs commentent le « surprenant » résultat du premier tour de l'élection présidentielle du 21 avril : le second tour opposera Chirac à Le Pen et non Chirac à Jospin comme l'annonçaient les sondeurs. Les sondeurs se sont encore trompés. On notera que les seuls sondages dont on peut vérifier la fiabilité sont les sondages préélectoraux, à condition qu'entre la date du sondage et celle du vote il n'y ait pas d'évolution de l'état de l'opinion publique et que les intentions de vote se traduisent par des votes réels. Il m'a semblé que le lecteur non averti retiendrait plus facilement les justifications de Roland Cayrol que les critiques de Michel Lejeune. Une formation du citoyen aux sondages passe par la compréhension des arguments exposés dans les deux articles, c'est l'objet de cet exposé.

Avant de poursuivre, précisons le vocabulaire utilisé. En France, le mot « sondage » désigne à la fois « l'enquête par sondage » (« *sampling* » en anglais) et le « sondage d'opinion » (« *poll* » en anglais). Le premier s'appuie sur une théorie probabiliste, le second est purement empirique. Nous allons parler aujourd'hui de « théorie et pratique des sondages » dans le premier sens du terme, c'est-à-dire *échantillonnage et estimation en populations finies*. Les deux sens du terme « sondage » se rejoignent dans la mesure où les sondeurs qui effectuent « les sondages d'opinion » par la méthode des quotas se réfèrent à la « théorie des sondages » pour justifier leurs pratiques.

Signalons une autre confusion dans l'utilisation des mots « hasard » et « aléatoire ». Dans le langage courant, un résultat est « aléatoire » ou un événement arrive « par hasard » lorsqu'il est imprévu, inattendu, subi. En probabilités et

(*) IUFM de Toulouse.
(**) Inspecteur général honoraire.

statistique, en revanche, les variables aléatoires et leur loi de probabilité sont des objets bien identifiés et le « hasard » est construit selon une loi de probabilité. En théorie des sondages, tirer un échantillon aléatoire (ou probabiliste) c'est extraire un échantillon de la population selon une loi de probabilité, sur l'ensemble des échantillons, que l'on s'est fixée à l'avance (cette phase est appelée « plan de sondage »). En particulier, tirer un échantillon aléatoire simple à probabilités égales de taille n (on dit abusivement au lycée « tirer au hasard ») signifie que l'on effectue cette opération de telle façon que tous les échantillons de taille n aient la même probabilité d'être tirés. Ceci n'est possible qu'en suivant des *procédures aléatoires* très contrôlées, par exemple, en reportant dans une liste les identifiants de chacun des individus de la population (liste appelée « base de sondage ») et en utilisant de façon adéquate le générateur de nombres « pseudo-aléatoires » de sa calculatrice ou de son ordinateur pour le tirage des individus. Pour toute précision on se reportera à l'article suivant : « Parzysz, B. (2005). Quelques questions à propos des générateurs aléatoires. *Statistique au lycée* vol. 1 (coord. par Chaput, B. & Henry, M.), 181-199. Éd. APMEP ». Un tel échantillon n'a donc rien à voir avec un échantillon d'individus rencontrés « par hasard » dans la rue.

Pour le statisticien, « sondage » (enquête auprès d'un échantillon de la population) s'oppose à « recensement » (enquête auprès de toute la population). À propos de recensement, il est intéressant de savoir que, jusqu'en 2004, le « dénombrement » de la population française se faisait par « recensement » tous les sept à neuf ans ; il se fait depuis par « sondage aléatoire » et par rotation tous les ans (cf. la présentation de la nouvelle méthode sur le site de l'INSEE : http://www.insee.fr/fr/nom_def_met/sources/sou-rp.htm) ; ce nouveau dénombrement permettra d'obtenir à partir de 2008 une amélioration sensible de la qualité de l'information.

Remarque : quelles sont les conditions préalables pour un sondage de qualité ?

Quand un sociologue (ou un politologue) commande (ou analyse) un sondage d'opinion, c'est pour recueillir des informations sur un état de celle-ci. Leur pertinence dépend de plusieurs facteurs :

1/ la qualité des questions posées,
2/ la qualité des réponses des sondés,
3/ l'utilisation de techniques statistiques appropriées,
4/ l'intégration éventuelle d'informations provenant d'autres sources.

Le présent exposé traite essentiellement du point 3/, mais examinons rapidement les deux premiers. Dans les études sur des élections à venir, les questions posées sont simples et ne comportent pratiquement pas d'ambiguïté. Il n'en est pas de même dans d'autres études. Par exemple pour la même question posée sous forme affirmative ou sous forme négative, les réponses sont très différentes. Soit une maladie grave et un traitement chirurgical lourd, si vous demandez : « le traitement réussit à 70%,vous feriez-vous opérer ? », alors une majorité répond oui. Si vous formulez la même question sous la forme suivante : « le traitement a un risque d'échec de 30%, vous feriez-vous opérer ? », alors une majorité répond non !

Il faut ensuite analyser la façon dont les individus de l'échantillon répondent. Il y a d'abord à s'interroger sur la sincérité des réponses. Ainsi, dans les sondages pré-

électoraux les intentions de votes pour les extrêmes sont souvent sous-estimées. Cela a été longtemps vrai pour le vote communiste, c'est probablement encore vrai pour le vote Front National, d'où les redressements effectués par les instituts spécialisés à partir des écarts observés à une élection précédente entre les intentions et le vote réellement enregistré. Mais la méthode est hasardeuse (sans jeu de mots). En 1981 les intentions de votes en faveur de Georges Marchais avaient été redressées à tort et le vote communiste surestimé : à cette date on ne cachait plus son intention de voter pour le candidat du parti communiste, mais le phénomène n'avait pas été détecté par les politologues.

La qualité des réponses dépend aussi de la réceptivité des individus interrogés. Si le sondé n'a pas réfléchi au préalable à la question posée il risque de répondre à peu près n'importe quoi. Que répondrait un intellectuel si on le mettait dans la situation de choisir entre travailler comme outilleur ou comme chaudronnier dans une usine ? Il faut également tenir compte de la disponibilité du répondant, de son état de fatigue physique ou psychologique, de la cohérence de ses choix. Des réponses à deux questions voisines peuvent être très différentes, ce qui causera de grandes difficultés d'interprétation.

Les questions précédentes ne relèvent pas de la statistique mais de la discipline qui a commandité le sondage, de la psychosociologie, même si des méthodes statistiques permettent de contourner certains obstacles, comme par exemple la méthode « Warner » pour inciter à donner des réponses sincères à des questions délicates. Cf. la méthode utilisée pour l'enquête auprès des lycéens sur l'usage de la drogue explicitée ci-dessous.

2. Marge d'erreur de 3% du sondage par quotas ?

Dans le premier article, Roland Cayrol considère que, pour cette élection, les critiques faites aux sondeurs sont infondées. Il présente un tableau montrant les résultats du dernier sondage (17/18 avril) et les résultats du premier tour de l'élection présidentielle du 21 avril : *l'écart en valeur absolue est inférieur à 3% pour chacun des candidats, marge d'erreur de la technique*. On a en effet, pour les trois premiers candidats :

	Sondage	Élection	Écart
Chirac	19.5%	19.7%	0.2%
Le Pen	14.0%	16.9%	2.9%
Jospin	18.0%	16.1%	1.9%

Imaginons que les résultats soient annoncés avec une marge d'erreur de 3%. Pour cela on construit un intervalle de confiance, ce qui signifie que la méthode utilisée pour l'établir avait une probabilité de 0,97 de recouvrir la vraie valeur inconnue de la proportion de votes pour tel ou tel candidat. Il est donc possible que cette dernière soit en dehors de l'intervalle de confiance, mais la probabilité d'un tel événement est faible. On obtient les résultats suivants :

Chirac	entre 16.5% et 22.5% des voix
Jospin	entre 15.0% et 21.0% des voix
Le Pen	entre 11.0% et 17.0% des voix

Il apparaît que Le Pen peut être second … et même premier ; toutes les configurations de l'ordre des trois premiers candidats étaient possibles.

Il est bien évident que si les sondeurs et les journalistes annonçaient que leur marge d'erreur est de 3%, nous ne serions pas inondés de sondages comme c'est le cas aujourd'hui.

3. Sondage aléatoire simple à probabilités égales ; marge d'erreur

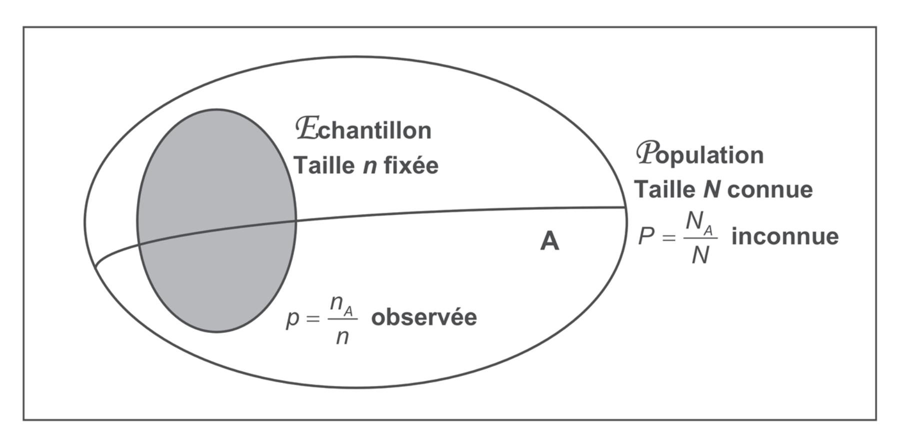
Voici à présent le premier extrait de l'article de Michel Lejeune.

Les derniers sondages indiquaient 18% pour Jospin et 14% pour Le Pen. Pour les rares scientifiques qui savent comment sont produites les estimations, l'écart rendait tout à fait plausible le scénario qui s'est réalisé. Si l'on se réfère à un sondage qui serait effectué dans des conditions idéales (tirage aléatoire absolu, taux de réponse 100%, aucune fausse déclaration) on obtient sur de tels pourcentages une incertitude de plus ou moins 3% étant donné la taille de l'échantillon.

Il est fait mention du premier résultat de la théorie des sondages : estimation par intervalle à 95% de confiance d'une proportion (sondage aléatoire simple de taille n à probabilités égales).

En voici une visualisation puis l'énoncé du théorème :

On note A une partie de la population, ce peut être, par exemple, la population qui vote « oui » à un référendum. La proportion $\mathrm{P} = \frac{\mathrm{N_A}}{\mathrm{N}}$ de A dans la population, avec des notations évidentes, est inconnue et l'objectif est d'estimer cette proportion (appelée paramètre d'intérêt) à partir d'une enquête par sondage. Après enquête auprès d'un échantillon de taille n, on peut calculer la proportion de A dans l'échantillon $p = \frac{n_\mathrm{A}}{n}$ (les notations sont encore évidentes).



Théorème (approché)

Soit A une partie de la population et P la proportion de A dans la population. Si l'on tire dans la population un échantillon de n individus selon une procédure aléatoire

garantissant l'égale probabilité de tirage des échantillons et si l'on observe une proportion p de A dans l'échantillon, alors, avec une probabilité de se tromper de 5% (c'est-à-dire, une confiance de 95%), la proportion P inconnue sur la population appartient à l'intervalle :

$$\left[p-2\sqrt{\frac{p(1-p)}{n}}\,;p+2\sqrt{\frac{p(1-p)}{n}}\right].$$

Dans l'énoncé du théorème on a utilisé le théorème central limite : asymptotiquement la loi binomiale que suit la variable nombre d'individus de A dans l'échantillon est proche de la loi de Gauss dite aussi loi normale.

On suppose ici que le *taux de sondage* n/N est « négligeable » (inférieur à 1/10), ce qui revient à assimiler le sondage sans remise (on extrait une partie de taille n de la population) à un sondage avec remise (on extrait un élément de la population n fois de suite dans les mêmes conditions).

L'idée de la preuve est la suivante. Reprenons l'exemple du référendum. Nous disposons d'une urne avec des millions de bulletins, dont une proportion P indique le « oui », le reste indiquant le « non ». Si l'on tire « au hasard » (c'est-à-dire avec équiprobabilité) un seul bulletin de l'urne, il indiquera « oui » ou « non » avec probabilités P et 1 − P respectivement (en appliquant la règle « nombre de cas favorables » sur « nombre de cas possibles »).

Si l'on répète l'expérience de tirer un bulletin de l'urne n fois dans les mêmes conditions (tirage avec remise d'un échantillon aléatoire de taille n), alors le nombre de « oui » est l'observation d'une variable aléatoire binomiale de taille n et de paramètre P, donc de moyenne n P et d'écart-type $\sqrt{n\mathrm{P}(1-\mathrm{P})}$ et la fréquence p de « oui » est l'observation d'une variable aléatoire de moyenne P et d'écart-type $\sqrt{\mathrm{P}(1-\mathrm{P})/n}$.

Deux théorèmes de probabilité permettent de conclure : *la loi des grands nombres* (la probabilité que p s'écarte de P de plus qu'un $\varepsilon > 0$ arbitraire tend vers 0 lorsque n augmente indéfiniment) et *le théorème central limite* (p est l'observation d'une variable aléatoire dont la loi est proche de la loi normale de moyenne P et d'écart type $\sqrt{\mathrm{P}(1-\mathrm{P})/n}$ pour n assez grand). Ce théorème permet de contrôler la vitesse de convergence de p vers P ; en particulier, plus de 95% des valeurs de p sont comprises entre $\mathrm{P}-2\sqrt{\mathrm{P}(1-\mathrm{P})/n}$ et $\mathrm{P}+2\sqrt{\mathrm{P}(1-\mathrm{P})/n}$.

La statistique inférentielle classique repose sur ces deux théorèmes. **C'est parce que l'on contrôle le comportement des observations faites sur des échantillons aléatoires que l'on peut donner des informations sur les paramètres inconnus de la population dont sont extraits les échantillons (en contrôlant la probabilité de se tromper).**

En particulier, on déduit du résultat précédent que 95% des échantillons aléatoires de taille n permettent de construire un intervalle

$$\left[p-2\sqrt{p(1-p)/n}\,;p+2\sqrt{p(1-p)/n}\right]$$

contenant la proportion inconnue P (estimation de P par intervalle à 95% de confiance). En moyenne, dans 95% des cas, la proportion P est donc dans l'intervalle indiqué. On peut espérer que c'est le cas pour l'échantillon aléatoire de taille n considéré.

La *marge d'erreur* à 95% de confiance est donnée dans le tableau suivant pour quelques valeurs de la proportion p observée sur l'échantillon et de la taille n de l'échantillon :

Marge d'erreur à 95% de confiance : $2\sqrt{\dfrac{p(1-p)}{n}}$

Proportion observée p / Taille échantillon n	10% ou 90%	20% ou 80%	30% ou 70%	40% ou 60%	50%
100	6.00%	8.00%	9.17%	9.80%	10.00%
400	3.00%	4.00%	4.58%	4.90%	5.00%
1 000	1.90%	2.53%	2.90%	3.10%	3.16%
5 000	0.85%	1.13%	1.30%	1.39%	1.41%
10 000	0.60%	0.80%	0.92%	0.98%	1.00%

On remarque que, à taille fixée de l'échantillon, c'est pour une proportion p proche de 50% (dernière colonne du tableau) que l'intervalle est le plus grand. Il s'écrit alors :

$$\left[p-\frac{1}{\sqrt{n}}\,;p+\frac{1}{\sqrt{n}}\right]$$

(les lecteurs avisés auront remarqué que le produit $p\ (1-p)$ est maximum si $p = 1 - p = 1/2$).

Pour un échantillon de taille 1 000, on a $1/\sqrt{1\,000}=3\%$; il s'agit de la *marge d'erreur* (qu'il serait en fait préférable d'appeler *marge d'incertitude*) indiquée dans l'article de Michel Lejeune mais aussi dans celui de Roland Cayrol. Cette marge d'erreur est parfois exprimée en « points » et non en «% » pour éviter de suggérer 3% de la proportion p, ce qui donnerait une marge erronée à 1.5% pour $p = 50\%$.

Si l'on observe une proportion p égale à 52% sur un échantillon de taille 1 000, au lieu de laisser croire que la proportion P inconnue est quasiment égale à 52% il faudrait annoncer que « la proportion P est comprise entre 49% et 55% » et annoncer de plus que cette affirmation n'est pas certaine, que cet intervalle a été construit avec un niveau de confiance de 95%. On peut donc se tromper dans 5% des cas.

Il est important de remarquer que la précision de l'estimation ne dépend pas de la taille N de la population et ne dépend pas non plus du taux de sondage s'il est inférieur à 1/10. Un sondage aléatoire simple à probabilités égales de taille 1 000 dans une

population de taille 20 000 000 (taux de sondage 1/20 000) est plus précis qu'un sondage aléatoire simple à probabilités égales de taille 600 dans une population de taille 12 000 (taux de sondage 1/20).

Illustrons ce résultat contre intuitif : pour savoir si la soupe est salée, une fois bien mélangée, il suffit de goûter une cuillerée de soupe, que cette cuillère soit extraite d'un petit bol ou d'un très grand chaudron.

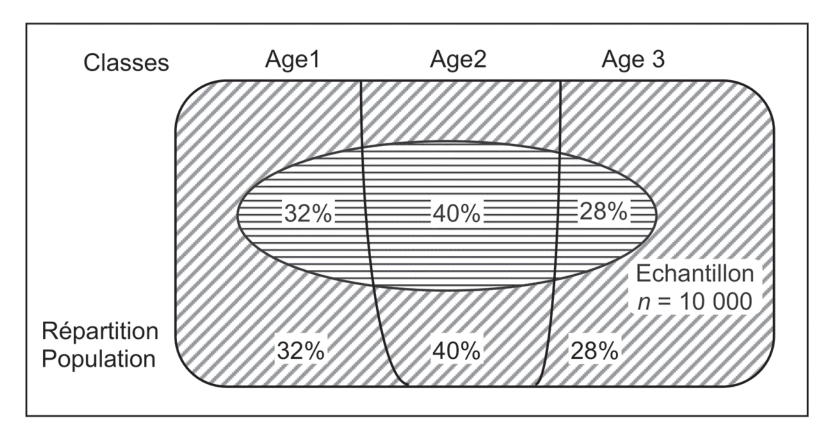
Enfin, ce résultat est valable dans les *conditions idéales (tirage aléatoire absolu, taux de réponse 100%, aucune fausse déclaration)* décrites par Michel Lejeune.

Ce résultat fait partie des thèmes d'études du programme de seconde. Le professeur qui le traite satisfait la curiosité de ses élèves de seconde, tout en concourant à leur formation citoyenne.

Remarque 1 : échantillon aléatoire simple « représentatif » pour *n* assez grand

Si la population est répartie selon trois classes d'âge selon les proportions (32%, 40%, 28%), alors un échantillon aléatoire simple à probabilités égales de taille 10 000 présente une répartition selon ces trois classes d'âges dans les mêmes proportions (à moins de 1% près). L'échantillon est quasiment un *modèle réduit* de la population pour ces trois classes d'âge. Lorsque l'on observe la même répartition dans l'échantillon et dans la population, on parle d'échantillon « *représentatif* », mais il faudrait préciser *par rapport à quel critère* ; ici il est quasiment (à moins de 1% près) représentatif par rapport aux trois classes d'âge.

	Âge 1	Âge 2	Âge 3
Échantillon Effectif : 10 000	32%	40%	28%
Population	32%	40%	28%



Le très grand avantage d'un échantillon aléatoire simple à probabilités égales de taille 10 000 est qu'*il est « représentatif »* (à 1% près) *par rapport à toutes les variables (ou critères) connues ou inconnues sur la population* … en particulier celle dont l'objectif est justement d'estimer « la répartition des votes selon les candidats à

une élection présidentielle » par exemple. Il permet donc d'estimer avec une très bonne précision n'importe quelle proportion ou répartition, *à condition qu'il n'y ait pas de non réponses ou de fausses déclarations*. Nous nous rapprochons de ces conditions lors des premières estimations faites à vingt heures à partir des premiers dépouillements, alors que les sondages pré-électoraux se limitant à des échantillons de taille 1 000, ne peuvent revendiquer cette représentativité (et surtout parce qu'il ne s'agit plus d'intentions, mais de vote réels).

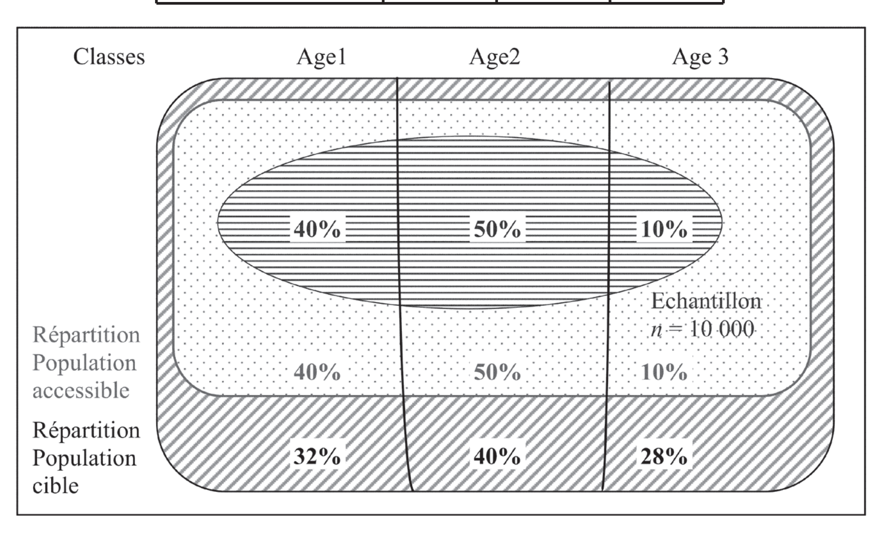
L'échantillon représentatif est l'objet du débat entre statisticiens qui eut lieu à la fin du XIXe siècle et au début du XXe siècle et qui se conclut dans les années trente sur l'indiscutable supériorité du « choix au hasard » sur le « choix d'experts » appelé aussi « choix raisonné ».

Remarque 2 : défaut de couverture, estimation biaisée

Bien sûr le résultat ci-dessus suppose que la base de sondage dans laquelle s'effectue le tirage de l'échantillon correspond bien à la liste des individus de la population ciblée par l'enquête. Lorsque la population accessible par les enquêteurs est différente de la population cible, on parle de *défaut de couverture*.

Supposons que l'on tire un échantillon de taille 10 000 dans une population accessible (par exemple, la population des internautes), sur laquelle la répartition des trois classes d'âge (dans l'ordre croissant) est (40%, 50%, 10%) et non (32%, 40%, 28%) comme c'est le cas dans la population cible (population des adultes français).

	Âge 1	Âge 2	Âge 3
Population cible	32%	40%	28%
Population accessible	40%	50%	10%
Échantillon Effectif 10 000	40%	50%	10%



L’échantillon aléatoire simple à probabilités égales de taille 10 000 sera, pour la répartition selon les trois classes d’âge, « représentatif » de la population accessible et non de la population cible. En résumé, quand on parle d’un « *échantillon représentatif* » il faut préciser *de quelle population et par rapport à quels critères*. Ici, comme il s’agit d’un échantillon aléatoire simple de taille 10 000 il sera « représentatif », par rapport à tous les critères, de la population accessible dans lequel il a été tiré.

Ce sera le cas également pour la répartition des votes selon les candidats. Si 46% de la population des internautes vote pour le candidat X alors que c’est le cas de 52% de la population cible, notre échantillon fournira la proportion 46% et non pas la proportion 52%, d’où une estimation biaisée de la proportion cherchée. Un défaut de couverture peut entraîner un biais important dans les estimations.

Remarque 3 : Redressement de l’échantillon

Bien sûr, il est facile de demander à chaque internaute enquêté de dire dans quelle classe d’âge il se situe. Grâce à l’INSEE, on connaît la répartition de la population cible selon les trois classes d’âge (32%, 40%, 28%), alors que notre échantillon de taille 10 000, fidèle « représentant » de la population accessible, se répartit selon les proportions (40%, 50%, 10%).

Répartition selon trois classes d’âge

	Âge 1	Âge 2	Âge 3	Ens.
Population cible	32%	40%	28%	100%
Pop. access. et échantillon	40%	50%	10%	100%

Supposons que les proportions d’intentions de vote pour le candidat X soient les suivantes :

Proportions de votants pour X

	Âge 1	Âge 2	Âge 3	Ens.
Population cible	50%	51%	56%	52%
Pop. access. et échantillon	45%	46%	54%	46%

La proportion globale de votants pour X observée sur l’échantillon, 46%, est une moyenne pondérée des proportions de votants pour X de chaque classe d’âge, les poids correspondant à la répartition de l’échantillon selon l’âge :

$$0.45\times0.40+0.46\times0.50+0.54\times0.10 \simeq 0.46.$$

Le *redressement d’échantillon* consiste à utiliser la répartition de la population cible (connue de façon précise par ailleurs) au lieu de la répartition de l’échantillon ; on obtient :

$$0.45\times0.32+0.46\times0.40+0.54\times0.28 \simeq 0.48.$$

On n’obtient toujours pas la proportion cherchée 52%.

Quelle que soit la classe d’âge, la proportion de ceux qui votent pour X parmi les internautes est différente de la proportion de ceux qui votent pour X dans la population cible. La différence entre 52% (proportion cherchée) et 46% (estimation biaisée) ne

s'explique pas uniquement par la différence de répartition, selon les classes d'âge, de la population des internautes et de la population cible.

Le redressement d'échantillon revient à donner un poids de 2.8 (= 0.28/0.10) à chaque internaute de la troisième classe d'âge, sous-représentée dans l'échantillon par rapport à la population cible ; de même, on donne un poids 0.8 (= 0.32/0.40) à chaque internaute de la première classe et un poids 0.8 (= 0.40/0.50) à chaque internaute de la deuxième classe. Cela permet de reconstituer un échantillon d'internautes « représentatif » de la population cible selon les classes d'âge, mais si les internautes ne votent pas comme la population cible, on aura toujours une estimation biaisée pour la proportion cherchée.

Remarque 4 : les non-réponses

Dans le cadre d'un sondage aléatoire simple, les individus de la population sont identifiés et ce sont les individus tirés par la procédure aléatoire de constitution de l'échantillon qui doivent répondre à l'enquête … et non pas d'autres individus !!

Même en supposant qu'ils soient tous joignables, il n'est pas rare que certains refusent de répondre à des questions qu'ils jugent sensibles. C'est un des problèmes les plus importants rencontrés dans la pratique des sondages car, bien souvent, le fait de répondre ou de ne pas répondre dépend justement de l'attitude par rapport à la question sensible que l'on cherche à mesurer.

Prenons la situation suivante, inspirée d'un exemple présenté par J.-C. Deville lors des dernières Journées de Statistique de la SFdS (Paris, 2006).

On demande aux 600 élèves d'un lycée (300 filles et 300 garçons) s'ils ont déjà consommé de la drogue. Les données, présentées dans le tableau suivant, sont fictives et, pour simplifier, concernent l'ensemble des élèves d'un lycée et non un échantillon.

Résultats de l'enquête sur les 600 élèves

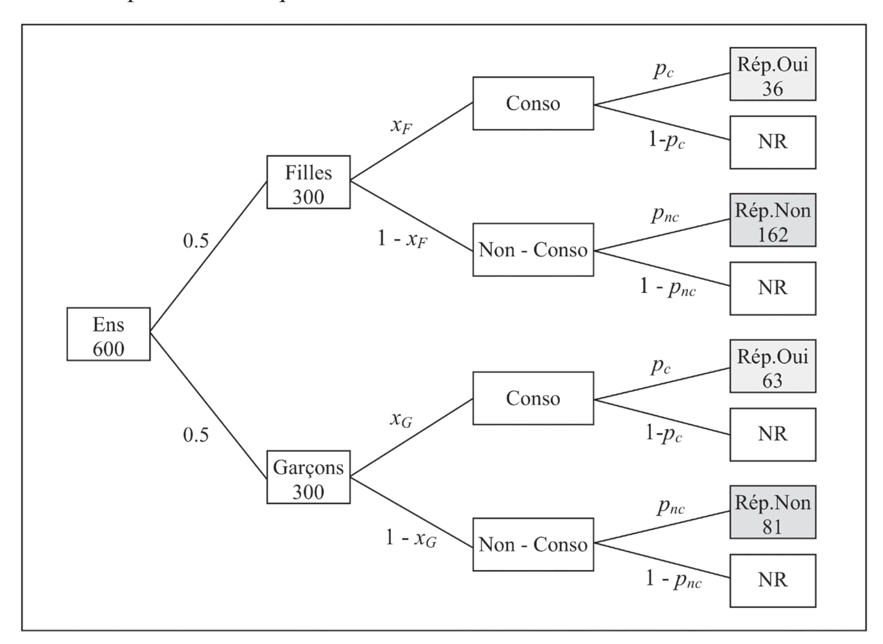
Consommation / Sexe	oui	non	Rép	% de oui	Non Rép	Ensemble
Filles	36	162	198	18%	102	300
Garçons	63	81	144	44%	156	300
Ensemble	99	243	342	29%	258	600

Si l'on suppose que ceux qui ont répondu n'ont pas fait de fausse déclaration, les « oui » correspondent à ceux qui ont déjà consommé de la drogue parmi les répondants. La proportion d'élèves qui déclarent avoir déjà consommé de la drogue est donc de 29% (18% de filles et 44% de garçons).

Mais nous cherchons à connaître la proportion d'élèves qui ont consommé de la drogue parmi les élèves du lycée et non parmi les élèves qui ont répondu : notons x_F et x_G ces proportions inconnues pour les filles et pour les garçons.

Supposons que le fait de répondre dépende du fait d'avoir consommé ou non de la drogue mais ne dépende pas du sexe ; on note p_c la proportion des répondants parmi ceux qui ont consommé et p_{nc} parmi ceux qui n'ont pas consommé ; ces proportions sont supposées identiques pour les filles et pour les garçons (modélisation du comportement de non réponse avec deux paramètres).

Nous pouvons alors représenter le problème par l'arbre de fréquence conditionnelle suivant et répondre à nos questions.



Transformons cet arbre en tableau :

Filles : 300

	Consommateurs : x_F	Non consommateurs : $1 - x_F$
Non réponses	$1 - p_c$	$1 - p_{nc}$
Réponses OUI	p_c (nb rép : 36)	
Réponses NON		p_{nc} (nb rép. : 162)

Garçons : 300

	Consommateurs : x_G	Non consommateurs : $1 - x_G$
Non réponses	$1 - p_c$	$1 - p_{nc}$
Réponses OUI	p_c (nb rép : 63)	
Réponses NON		p_{nc} (nb rép. : 81)

On a alors :

$$x_G p_c = 63/300, (1-x_G)p_{nc} = 81/300\,; x_F p_c = 36/300, (1-x_F)p_{nc} = 162/300.$$

Des calculs simples permettent d'obtenir :

$$p_c = 0.3,\ p_{nc} = 0.9,\ x_F = 0.4,\ x_G = 0.7.$$

On obtient une proportion de répondants de 30% parmi ceux qui ont consommé et de 90% parmi ceux qui n'ont pas consommé ; 40% des filles du lycée et 70% des garçons du lycée ont déjà consommé de la drogue.

On en déduit que 55% (= (0.4 + 0.7)/2) des élèves du lycée ont déjà consommé de la drogue et non pas 29% comme annoncé au départ en ne tenant compte que de ceux qui ont répondu.

Remarquons que l'ensemble des répondants joue un rôle analogue à celui des internautes de l'exemple précédent. La réponse à notre question de consommation de drogue est connue sur l'ensemble des élèves qui ont accepté de répondre (population disponible) et non sur l'ensemble des élèves du lycée (population cible). Cette analogie est la raison pour laquelle le problème des non réponses est présenté dans ce paragraphe.

Pour des questions sensibles susceptibles d'engendrer une forte proportion de non-réponses, il est possible, pour éviter les non-réponses, de mettre en place un dispositif astucieux reposant sur une procédure aléatoire de recueil des réponses (cf. annexe 1).

Remarque 5 : estimation d'une proportion dans une sous-population

Prenons l'exemple fictif du résultat annoncé d'une enquête sur les départs en congés par sondage aléatoire simple à probabilités égales avec remise, de taille 1 000 :

« 15% sont partis en vacances ; parmi eux, 52% sont partis à l'étranger ».

La marge d'erreur à 95% de confiance du premier résultat est de 2% $\left(=2\sqrt{\dfrac{0.15\times0.85}{1\,000}}\right)$ mais celle du deuxième résultat est de 8% $\left(=2\sqrt{\dfrac{0.52\times0.48}{150}}\simeq\dfrac{1}{\sqrt{150}}\right)$ car la taille de l'échantillon de ceux qui ont répondu à cette deuxième question est à présent de 150 et non de 1 000.

Dans les sondages préélectoraux des présidentielles, il est annoncé par exemple que, parmi les 15% de ceux qui ont l'intention de voter pour tel candidat, 52% reporteront leur voix sur tel autre candidat. La marge d'erreur sur la deuxième information est quatre fois plus importante que celle sur la première.

4. Échantillon aléatoire stratifié (proportionnel puis optimal)

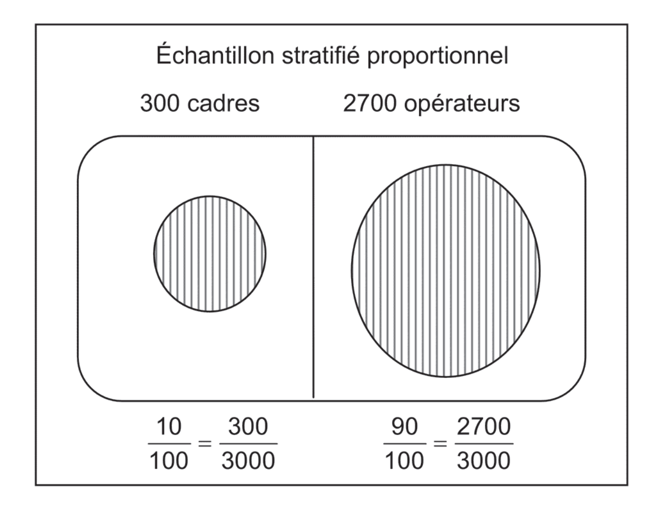
Prenons l'exemple suivant, pas très réaliste mais permettant bien d'illustrer la situation : l'ensemble des salariés d'une entreprise est composé de 300 cadres et 2 700 opérateurs. On cherche à estimer la masse salariale de l'année distribuée par cette entreprise à l'aide d'une enquête auprès d'un échantillon de 100 salariés à qui on demande de bien vouloir donner leur salaire mensuel moyen net.

Échantillon aléatoire stratifié proportionnel

Dans la liste des noms des 3 000 salariés, est indiqué le statut (cadre ou opérateur) (ce type d'information est appelée « information auxiliaire »).

Plutôt que tirer un échantillon aléatoire simple de taille 100 dans l'ensemble des 3 000 individus, on peut tirer deux échantillons aléatoires simples de tailles respectives 10 et 90 dans l'ensemble des cadres et dans l'ensemble des opérateurs. Il s'agit alors d'un échantillon aléatoire stratifié (les cadres et les opérateurs constituent les deux strates) représentatif de l'ensemble des salariés pour le statut puisque la répartition des cadres et des opérateurs dans l'échantillon est la même que dans la population. Plutôt que « représentatif » on utilisera l'adjectif « proportionnel ». On a en effet proportionnalité entre les tailles des échantillons des strates et les tailles des strates : $n_k \propto \mathrm{N}_k$.

On montre que, à taille d'échantillon constante, la précision des estimations peut être considérablement améliorée en utilisant l'échantillon aléatoire stratifié proportionnel plutôt que l'échantillon aléatoire simple ; l'amélioration est d'autant meilleure que les moyennes par strates de la variable objet de l'enquête (ici le salaire) sont différentes (cf. annexe 2).



Échantillon aléatoire stratifié optimal

Intuitivement, cette répartition proportionnelle de l'échantillon ne semble pas la meilleure. En effet, les opérateurs ont quasiment tous les mêmes salaires assez proches du SMIC alors que les cadres ont des salaires bien plus dispersés entre les jeunes ingénieurs et les dirigeants plus âgés. Interroger 90 personnes qui donnent quasiment la même réponse et seulement 10 cadres pour estimer le salaire moyen des cadres ne semble pas idéal.

On montre en effet que, à taille d'échantillon global fixé, la répartition de l'échantillon entre opérateurs et cadres qui donne la meilleure précision n'est pas l'échantillon proportionnel mais un échantillon qui tient compte également de la

dispersion selon les strates de la variable objet de l'enquête. Plus précisément, la répartition optimale est celle dans laquelle les tailles des échantillons sont proportionnelles aux produits de la taille de la strate par l'écart-type du caractère dans la strate correspondante :

$$n_k \propto \mathrm{N}_k \, \sigma_k.$$

L'amélioration de la répartition optimale sur la répartition proportionnelle est d'autant meilleure que les écarts-types par strates sont différents (cf. annexe 2).

L'échantillon aléatoire stratifié proportionnel n'est donc pas, pour le statisticien spécialiste des sondages, la panacée. Tout son travail repose sur la recherche du meilleur échantillon aléatoire en fonction de l'objectif de l'étude et des moyens (budget, faisabilité) disponibles. Il utilise les informations auxiliaires pour construire le plan d'échantillonnage et évaluer les précisions des estimateurs associés.

Tous les échantillons aléatoires dont toutes les étapes, de la procédure de sélection jusqu'au suivi sur le terrain, sont dûment contrôlées, sont alors considérés comme des échantillons « représentatifs ». En effet, grâce au calcul des probabilités, on peut inférer, à partir des résultats obtenus sur l'échantillon, les résultats concernant la population cible et donner une estimation par intervalle de confiance des paramètres de la population.

L'échantillon stratifié est un échantillon à deux degrés : on répartit la population en strates (unités primaires) et on tire un échantillon d'individus (unités secondaires) dans chaque strate. On fait donc un recensement au premier niveau. On peut utiliser plusieurs degrés. Par exemple, l'ensemble des ménages français est réparti selon les catégories d'agglomération de leur lieu d'habitation : rural, agglomération de moins de 20 000 habitants, agglomération de 20 000 à 100 000 habitants, agglomération de plus de 100 000 habitants, agglomération parisienne, puis selon les cantons à l'intérieur des catégories. On tire un échantillon aléatoire de cantons (unités secondaires) dans chaque catégorie d'agglomération (strates au premier degré). Pour chaque canton de l'échantillon, on utilise une nouvelle partition géographique en quartiers (unités tertiaires) et on tire un échantillon de quartiers. On constitue alors la base de sondage des ménages des quartiers sélectionnés avant de tirer un échantillon aléatoire de ménages (unités quaternaires).

5. Méthode empirique des quotas

Reprenons l'article de Michel Lejeune :

En pratique, on est loin de se trouver dans ces conditions : on ne peut pas réaliser un véritable tirage aléatoire parmi les électeurs, le taux de réponse est de 10 à 20% en pratique dans le cas du téléphone, enfin, les fausses déclarations d'intention ne sont pas négligeables, en particulier sur les intentions de vote pour l'extrême droite. Les sondeurs croient, ou feignent de croire, que grâce à l'utilisation de quotas et au redressement des échantillons, leur précision pourrait être meilleure que, par exemple, ces 3%.

En effet, les sondeurs utilisent généralement la méthode empirique des quotas. Il ne s'agit pas d'une méthode aléatoire. L'unique contrainte est que l'échantillon

respecte les répartitions, connues sur la population, de quelques variables (dites de quotas). Constituons, par exemple, un échantillon d'adultes français de taille 1 000, « représentatif » de la population française par rapport au sexe, classes d'âge et PCS du chef de ménage :

		POPULATION	ÉCHANTILLON
SEXE	Hommes	48%	480
	Femmes	52%	520
	Ensemble	*100%*	*1 000*
ÂGE	< 35	27%	270
	35 à 65	51%	510
	>65	22%	220
	Ensemble	*100%*	*1 000*
PCS	PCS +	33%	330
	PCS −	30%	300
	Inactif	37%	370
	Ensemble	*100%*	*1 000*

Un groupe d'enquêteurs dans le central téléphonique de l'institut de sondage compose alors des milliers de numéros de téléphones et demande aux personnes qui acceptent de répondre (en plus des questions qui font l'objet de l'enquête) quel est leur sexe, leur âge et la PCS du chef de ménage jusqu'à boucler les quotas, c'est-à-dire jusqu'à obtenir un échantillon de taille 1 000 composé comme indiqué dans le tableau. Les quotas sont parfois donnés en fonction des effectifs croisants sexe et âge.

Seulement 10% à 20% des personnes contactées par téléphone acceptent de répondre à ce type d'enquête… À condition que la base de numéros de téléphone couvre bien l'ensemble des français ayant un téléphone fixe, la population disponible est donc « l'ensemble des personnes qui acceptent de répondre à une enquête téléphonique sur leur téléphone fixe ».

Inutile d'ajouter, après les développements que nous avons vus précédemment, que le défaut de couverture et le biais peuvent être très importants.

Les exemples de biais dus à la méthode sont nombreux. Une enquête par quotas (sexe, âge, PCS) auprès de la population toulousaine sur l'utilisation des transports en commun, entièrement réalisée en centre ville, a donné des résultats inutilisables, car pratiquement tous les individus de l'échantillon habitaient le centre et ne prenaient pas les transports en commun. Il est évident qu'il aurait fallu prendre en compte dans les quotas la répartition de la population selon les quartiers de la ville.

Autre exemple : une enquête sur la santé, réalisée par quotas auprès de la population de plus de 15 ans d'une ville, révèle une proportion anormalement élevée de personnes malades : l'enquête avait eu lieu à domicile pendant les heures de bureau !

Le grand avantage de la méthode des quotas est qu'elle ne nécessite pas de disposer d'une base de sondage d'où, comparativement à un sondage aléatoire de même taille, un très faible coût et une très grande rapidité. **L'inconvénient est qu'il n'est pas possible de calculer la précision des estimations obtenues**.

La méthode des quotas est une imitation d'un sondage aléatoire stratifié proportionnel, qui, on l'a vu, est meilleur qu'un sondage aléatoire simple. C'est la raison pour laquelle les sondeurs *croient* (*ou feignent de croire*), comme l'écrit Michel Lejeune, que la méthode des quotas est meilleure que le sondage aléatoire simple. Ils annoncent (quand ils le font) la marge d'erreur d'un échantillon aléatoire simple de même taille ; cela explique la marge de 3% du sondage par quotas de taille 1 000 annoncée par Roland Cayrol.

Dans le livre de Roland Cayrol intitulé « Sondages, mode d'emploi », on lit, page 38 :

C'est le fameux exemple des boules rouges et des boules noires, tirées dans un sac qui comporte le même nombre de boules rouges que de boules noires.

Si l'on tire un échantillon de 100 boules du sac – mais attention, d'une manière purement et absolument aléatoire –, la théorie des probabilités indique qu'il y a 95 chances sur 100 (attention encore : pas cent pour cent) pour que le nombre de boules rouges ainsi tirées au hasard soit compris entre 45 et 55. Nous savons qu'il y a 50% de boules rouges dans le sac ; donc la proportion exacte de boules rouges aura été approchée, par le tirage d'un échantillon aléatoire, avec une marge d'erreur, un intervalle de confiance, de 10% (45 à 55). Et il reste encore cinq chances sur cent que le résultat du tirage soit extérieur à cette fourchette.

Il est vrai que la marge d'erreur à 95% est de 10% (cf. le tableau présenté précédemment), ce qui donne un intervalle de confiance de la forme 50% ± 10% et non 50% ± 5% comme l'écrit Roland Cayrol ! La confusion vient peut-être de l'écriture ambiguë signalée plus haut : 10% de 50% correspond bien à 5%. On a pris des pourcentages relatifs pour des pourcentages absolus.

Ces quelques lignes sont la seule partie « technique » du livre de Roland Cayrol (130 pages), ce qui prouve que la pratique des sondages d'opinion n'a pas grand-chose à voir avec la théorie des sondages.

Les sondeurs s'abritent parfois derrière le fait que l'INSEE utilise la méthode des quotas (contrairement aux instituts nationaux de statistique anglo-saxons), mais l'INSEE n'utilise les quotas qu'en dernier degré d'un sondage aléatoire à plusieurs degrés et *après avoir validé la méthode par un sondage aléatoire pur jusqu'au dernier degré*.

Une enquête par la méthode des quotas correctement menée (reposant sur des modèles de comportement de la population pour la définition des variables de quotas, sur une population accessible « représentative » de la population cible pour la variable d'intérêt, sur une taille d'échantillon importante, sur un taux de non-réponses faible, utilisée éventuellement en dernier degré d'une enquête, les autres étant aléatoires, …) peut donner de bons résultats. Des études théoriques sur la justification et les limites de la méthode ont été publiées (cf. en particulier, « *Une théorie des enquêtes par quotas* » de J.-C. Deville, Technique d'enquête, Statistique Canada, 17-2, 1991).

Nous critiquons ici l'utilisation qui en est faite par les sondeurs et journalistes en période préélectorale. L'argument selon lequel il s'agit seulement d'*une photographie* qui ne permet pas de *prédire* le résultat final car les intentions de vote sont fluctuantes

n'est pas le moindre irrecevable. En effet, même le jour de la publication, les commentaires sollicitent les données numériques bien au-delà des résultats du sondage, même si l'on tient compte de marges d'erreur … qu'il est impossible de mesurer !

Poursuivons l'article de Michel Lejeune :

Un sondeur a récemment déclaré que le vote Le Pen constaté dans les échantillons devait être multiplié (c'est-à-dire, redressé) par 2 … selon des règles mathématiques rigoureuses, ajoutait-il.

D'où vient ce coefficient 2 (confirmé par d'autres sources) ? … tout rond ! Pourquoi pas 2.1 ou 1.8 (en fait il aurait fallu prendre 2.4). Nous avons vu que le « redressement » d'échantillon aléatoire repose sur la connaissance d'une information précise connue sur la population, par exemple grâce à l'INSEE. Ce coefficient 2 serait obtenu en demandant aux personnes enquêtées pour qui elles ont voté précédemment et en comparant les résultats bruts du sondage à cette question avec les résultats réellement obtenus précédemment.

En fait les corrections relèvent plus de l'appréciation empirique des politologues que de la théorie statistique. J'en veux pour preuve la surprenante proximité des résultats d'un institut à l'autre. Les sept derniers sondages publiés donnaient tous Jospin à 18% : pour tout observateur avisé, cette constance est statistiquement invraisemblable avec des échantillons de taille 1 000.

Vérifions cette dernière affirmation. Les résultats sont donnés à 0.5% près. Pour un sondage aléatoire simple à probabilités égales de taille 1 000 extrait d'une population dans laquelle la catégorie d'intérêt représente 18%, calculons la probabilité que la proportion observée sur l'échantillon soit comprise entre 17.5% et 18.5%. La proportion p est l'observation d'une variable aléatoire qui suit une loi normale de moyenne 0.18 et d'écart-type 0.012 $\left(=\sqrt{0.18\,(1-0.18)\,/\,1\,000}\right)$; la probabilité cherchée est donc de 0.32. On en déduit que la probabilité que sept sondages de taille 1 000 réalisés indépendamment et dans les mêmes conditions donnent le même résultat 18% est égale à $(0.32)^7$ soit 3 pour 10 000 !!!

Il faut donc comprendre que ces résultats sont « moyennés » pour tenir compte de tous les sondages réalisés les derniers jours, ce qui explique que sondeurs et journalistes les commentent comme s'ils étaient quasi-certains (comme pour un échantillon aléatoire stratifié proportionnel de taille 7 000). De plus, ils sont « redressés », ce qui signifie qu'ils sont obtenus à partir de modèles qui ont normalement pour objectif de réduire les marges d'erreur.

Ce n'est qu'après les élections que les sondeurs rappellent qu'il y a une marge d'erreur de 3% inhérente à la méthode … que les journalistes oublieraient de mentionner.

Michel Lejeune ajoute :

Quant aux médias, ils sont les dupes, peut-être les complices, des discours lénifiants des instituts.

Ce sont principalement les médias qui commandent les sondages préélectoraux et qui les commentent. Avant de converger de façon suspecte, les médias publient pendant des mois des sondages aux résultats contradictoires pourtant réalisés les mêmes jours. Il n'y a qu'en France que sont publiés autant de sondages reposant sur une méthodologie aussi peu fiable.

6. Les sondages, délaissés par les statisticiens et malmenés par les politologues

Voici le dernier extrait de l'article de Michel Lejeune qui explique le titre de son article « *Fautes de contrôles* » et de cette intervention (et de ce paragraphe) :

Il faut savoir qu'il n'y a aucun statisticien spécialiste de la théorie des sondages dans les cinq instituts concernés. Aujourd'hui la communauté scientifique doit se sentir en partie responsable de ce qui vient de se produire. Par dédain, elle n'a jamais voulu s'intéresser à la pratique des sondages, ni vraiment d'ailleurs, à la théorie.

Effectivement, pendant des années, la pratique des sondages était considérée par les statisticiens universitaires comme une activité ne concernant que la statistique publique (INSEE, INED, …), ne présentant donc aucun intérêt pour la formation des étudiants, ni sur un plan pédagogique, ni pour une recherche théorique. Même à l'INSEE, les recherches étaient davantage tournées vers l'économétrie que vers la théorie de l'échantillonnage.

Encore aujourd'hui, certains cursus universitaires de statistique ne comportent ni cours sur les sondages, ni même sur les plans d'expériences, les deux domaines dans lesquels le recueil des données statistiques est construit en fonction de l'objectif de l'étude.

C'est en octobre 1986 que l'Association des Statisticiens Universitaires (ancêtre de la SFdS) organise les deuxièmes Journées d'Étude en Statistique sur « Les Sondages », journées de formation pour statisticiens. Depuis, des cours ont été créés dans plusieurs universités françaises, des thèses ont été soutenues et des colloques sur la théorie et la pratique des sondages sont régulièrement organisés ; ils permettent de rassembler statisticiens, sociologues, politologues, théoriciens ou praticiens.

C'est peut-être à partir de telles rencontres qu'il sera possible de rediscuter du contrôle de qualité des sondages électoraux dont une Commission des Sondages, créée en 1977, est chargée de vérifier le suivi !

LES « THÉMATIQUES » DE TANGENTE

Nous saluons cette nouvelle formule de quatre numéros annuels d'au moins 52 pages qui servent de base aux H.S. de la Bibliothèque Tangente co-diffusée par l'A.P.M.E.P.

Son numéro 32, sous la direction de Raphaël DOUADY, est un remarquable « **Mathématique et Finance** », en prise sur l'actualité ! L'ouvrage rêvé pour, sur ce sujet, tout « honnête homme » de notre temps…

Henri BAREIL

Annexe 1

Anonymat : réponse aléatoire lors d'une enquête

(à partir des documents d'accompagnement des nouveaux programmes de Terminales)

Dans le cadre d'une enquête sur le tabac dans un lycée, la question « *est-ce que vous fumez plus d'un paquet de cigarettes par semaine : oui ? non ?* » peut être considérée comme sensible, certains lycéens peuvent avoir des difficultés à répondre sincèrement.

Afin de protéger l'anonymat du lycéen quant à sa consommation de tabac, y compris vis-à-vis de l'enquêteur, l'enquêteur propose à chaque lycéen de réaliser en son absence la procédure suivante.

Lancer une pièce.
Si elle tombe sur pile, répondre à la question : « est-ce que vous fumez plus d'un paquet de cigarettes par semaine : oui ? non ? »
Si elle tombe sur face, relancer une deuxième fois la pièce et répondre à la question : « est-ce que vous êtes tombé sur pile au deuxième lancer : oui ? non ? »

Lorsque le questionnaire porte la mention « oui » (resp. « non ») l'enquêteur ne peut pas savoir si le lycéen a répondu à la première ou à la deuxième question.

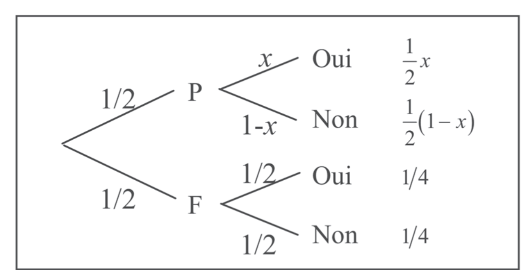
Soit p la proportion de « oui ». Proposer une modélisation de cette procédure aléatoire et calculer en fonction de p la proportion de fumeurs (de plus d'un paquet de cigarettes par semaine) du lycée.

Solution

Soit Ω la population de lycéens qui ont participé à l'enquête.
Soit x la proportion de fumeurs et p la proportion de ceux qui ont répondu « oui » à la question.
On suppose que 1 lycéen sur 2 a obtenu pile (resp. face) au premier lancer (resp. au deuxième lancer le cas échéant) et on suppose que le fait de fumer est indépendant du fait d'obtenir pile lors du lancer d'une pièce de monnaie.

On a alors l'arbre de fréquences conditionnelles suivant :

PILE (1/2)		FACE (1/2)	
OUI (x)	NON ($1-x$)	OUI (1/2)	NON (1/2)



On en déduit : $p=\frac{1}{2}x+\frac{1}{4}$ et donc $x=2p-\frac{1}{2}$. On peut donc estimer la proportion x de fumeurs chez les lycéens sans rien connaître des réponses individuelles.

Annexe 2

Sondage aléatoire simple et sondage aléatoire stratifié

Soit X une variable quantitative définie sur une population de taille finie N, de moyenne μ et d'écart-type σ. On cherche à estimer μ à partir des résultats obtenus sur un échantillon de taille n. Il est d'usage de noter $\hat{\mu}$ un estimateur de μ calculé à partir des observations faites sur l'échantillon. Un estimateur est une variable aléatoire définie sur l'ensemble des échantillons (c'est-à-dire, dont les observations varient d'un échantillon à un autre) et on cherche à ce qu'il soit « le plus proche » possible de μ. On choisira par exemple un critère de moindre carré, c'est-à-dire choisir $\hat{\mu}$ rendant minimum $\mathrm{E}\left(\left[\hat{\mu}-\mu\right]^2\right)$ (erreur quadratique moyenne). Comme on a : $\mathrm{E}\left(\left[\hat{\mu}-\mu\right]^2\right)=\mathrm{Var}\left(\hat{\mu}\right)+\left[\mathrm{E}\left(\hat{\mu}\right)-\mu\right]^2$ (variance plus carré du « biais »), le critère conduit souvent à des estimateurs sans biais $\left(\mathrm{E}\left(\hat{\mu}\right)=\mu\right)$ de variance minimale.

Un exemple illustrant les résultats qui suivent est traité à la fin de cette annexe.

Sondage aléatoire simple à probabilités égales avec remise

Soit $\bar{x}$ la moyenne observée sur un échantillon aléatoire simple de taille n à probabilités égales avec remise ; il s'agit de l'observation d'une variable aléatoire, notée $\overline{\mathrm{X}}$, d'espérance mathématique (ou moyenne) μ (donc $\overline{\mathrm{X}}$ est un estimateur sans biais de μ) et de variance σ^2/n. Si l'on note s^2 la variance corrigée de l'échantillon (c'est-à-dire la somme des carrés des écarts à la moyenne divisée par $n-1$ au lieu de n), alors s^2 est l'observation d'une variable aléatoire, notée S^2 d'espérance mathématique σ^2. On en déduit que $\hat{\mathrm{V}}\left(\overline{\mathrm{X}}\right)=\mathrm{S}^2/n$ est un estimateur sans biais de $\mathrm{V}\left(\overline{\mathrm{X}}\right)=\sigma^2/n$. Dès que la taille de l'échantillon est suffisamment grande (supérieure ou égale à 30), la loi de probabilité de $\overline{\mathrm{X}}$ peut être approchée par une loi gaussienne et on peut estimer μ par l'intervalle à 95% de confiance : $\bar{x}\pm 1.96\sqrt{\hat{\mathrm{V}}\left(\bar{x}\right)}$, c'est-à-dire $\bar{x}\pm 1.96\frac{s}{\sqrt{n}}$ (ici, $\hat{\mathrm{V}}\left(\overline{\mathrm{X}}\right)$, $\hat{\mathrm{V}}\left(\bar{x}\right)$ désignent une observation de la variable aléatoire notée de la même façon).

Sondage aléatoire stratifié à probabilités égales avec remise

La population est répartie en H sous-populations de tailles N_h ($h = 1,\ldots, H$) avec $N = \sum_{h=1}^{H} N_h$. On tire, de façon indépendante dans les H sous-populations (appelées strates), H échantillons aléatoires à probabilités égales avec remises, de tailles fixées n_h ($h = 1,\ldots, H$) et on pose $n = \sum_{h=1}^{H} n_h$. On note μ_h et σ_h la moyenne et l'écart-type de X sur la sous-population S_h ($h = 1,\ldots, H$).

On a alors les égalités (décomposition de la moyenne et de la variance sur la partition de la population que constituent les strates) :

$$\mu = \sum_{h=1}^{H} \frac{N_h}{N} \mu_h \text{ et } \sigma^2 = \sigma^2_{\text{intra}} + \sigma^2_{\text{inter}}$$

avec $\sigma^2_{\text{intra}} = \sum_{h=1}^{H} \frac{N_h}{N} \sigma_h^2$ (moyenne des variances) et $\sigma^2_{\text{inter}} = \sum_{h=1}^{H} \frac{N_h}{N} \left(\mu_h - \mu\right)^2$ (variance des moyennes).

On observe que la dispersion globale σ^2 est la somme de la dispersion à l'intérieur des classes et de la dispersion entre les classes. Cette dernière est d'autant plus grande que la division en classes est plus en relation avec le caractère étudié.

Si on note $\overline{x_h}$ la moyenne de la variable sur l'échantillon, c'est l'observation d'une variable aléatoire, notée $\overline{X_h}$, d'espérance mathématique μ_h et de variance σ_h^2 / n_h .

On estimera donc μ par la variable aléatoire $\overline{X_{St}} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{X_h}$ qui a pour espérance mathématique μ (donc $\overline{X_{St}}$ est un estimateur sans biais de μ) et pour variance $V\left(\overline{X_{St}}\right) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{\sigma_h^2}{n_h}$.

Si on note s_h^2 la variance corrigée de l'échantillon (observation d'une variable aléatoire notée S_h^2) alors la variance de $\overline{X_{St}}$ peut être estimée sans biais par $\hat{V}\left(\overline{X_{St}}\right) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{n_h}$ et, dès que les tailles de l'échantillon le permettent, la loi de probabilité de $\overline{X_{St}}$ peut être approchée par une loi normale de moyenne μ et de variance $\hat{V}\left(\overline{X_{St}}\right)$ (observation de la variable aléatoire notée de la même façon) d'où

une estimation de μ par l'intervalle à 95% de confiance $\overline{x_{St}} \pm 1.96\sqrt{\hat{V}\left(\overline{x_{St}}\right)}$.

Dans le cas particulier d'un *échantillon aléatoire stratifié proportionnel* défini par $n_h = \frac{N_h}{N} n$ (h = 1,…, H), on vérifie : $V\left(\overline{X_{St}}\right) = \frac{\sigma^2_{\text{intra}}}{n}$. Comme l'on a $\sigma^2_{\text{intra}} \leq \sigma^2$, avec égalité lorsque les moyennes des strates sont égales, l'intervalle de confiance est plus petit dans le cas du sondage stratifié proportionnel que dans le cas d'un sondage aléatoire simple de même taille n, et d'autant plus petit que la dispersion des moyennes des strates est grande.

Échantillon aléatoire stratifié optimal

On peut chercher, à taille fixée n, quelle est la répartition (on parle aussi d'allocation) $\left(n_1,\ldots,n_H\right)$ qui donne un estimateur de variance minimale. On montre que la fonction de $\left(n_1,\ldots,n_H\right)$ définie par : $V\left(\overline{X_{St}}\right) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 \frac{\sigma_h^2}{n_h}$, est minimale, sous la contrainte $\sum_{h=1}^{H} n_h = n$, pour $n_h = \frac{N_h \sigma_h}{\sum_{h=1}^{H} N_h \sigma_h}$ (h = 1,…, H). Cette répartition définit l'échantillon optimal et on obtient sous cette répartition $V\left(\overline{X_{St}}\right) = \frac{\bar{\sigma}}{n}$ avec $\bar{\sigma} = \sum_{h=1}^{H} \frac{N_h}{N}\sigma_h$ (moyenne des écarts-types).

Comme on a $\bar{\sigma} \leq \sigma$, avec égalité lorsque les écarts-types des strates sont égaux, l'intervalle de confiance est d'autant plus petit que les écarts-types des strates sont différents.

Remarques

Dans le cas où X est l'indicatrice d'une sous-population A (c'est-à-dire, égale à 1 pour un élément de A, égale à 0 sinon), soit P la proportion de A par rapport à la population. Alors la moyenne μ de X est P et l'écart-type σ de X est $\sqrt{P(1-P)}$. L'estimation d'une proportion est un cas particulier d'estimation d'une moyenne.

Dans le cas d'un échantillon aléatoire de taille n à probabilités égales *sans* remise, on montre que la moyenne d'échantillonnage $\bar{X}$ est un estimateur sans biais de μ de variance $V\left(\bar{X}\right) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{\sigma_c^2}{n}\left(1-\frac{n}{N}\right)$ en posant $\sigma_c^2 = \sigma^2 \frac{N}{N-1}$ et que la variance corrigée de l'échantillon S^2 est un estimateur sans biais de σ_c^2 ; on en déduit que

$\hat{V}(\bar{X}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$ est un estimateur sans biais de $V(\bar{X})$. Les intervalles de confiance sont donc réduits par rapport à un échantillon aléatoire avec remise, la réduction étant d'autant plus importante que le taux de sondage n/N est proche de 1. On retrouve le résultat intuitif selon lequel la variance de la moyenne d'échantillonnage $\bar{X}$ est nulle dans le cas d'un recensement (n = N).

Application

Reprenons l'exemple proposé dans le texte. L'ensemble des salariés d'une entreprise est composé de 300 cadres et 2 700 opérateurs. Dans chacun des deux groupes, la moyenne et l'écart-type des salaires sont donnés dans le tableau suivant :

	Effectif	Salaire moyen	Écart-type
Cadres	300	3 200	1 200
Opérateurs	2 700	1 500	100
Ensemble	3 000	?	?

Décomposition de la moyenne et de la variance sur la partition « cadres/opérateurs »

Soit μ et σ la moyenne et l'écart-type de la distribution des salaires sur l'ensemble des salariés. Alors on a :

$\mu = \frac{1}{10} \times 3\,200 + \frac{9}{10} \times 1\,500 = 1\,670$ (la moyenne de l'ensemble est la moyenne des moyennes) ;

$\sigma^2_{\text{intra}} = \frac{1}{10} \times (1\,200)^2 + \frac{9}{10} \times (100)^2 = 153\,000 \simeq (391)^2$ (la variance intra est la moyenne des variances des groupes) ;

$\sigma^2_{\text{inter}} = \frac{1}{10} \times (3\,200 - 1\,670)^2 + \frac{9}{10} \times (1\,500 - 1\,670)^2 = 260\,100$ (la variance inter est la variance des moyennes des groupes) ;

$\sigma^2 = \sigma^2_{\text{inter}} + \sigma^2_{\text{intra}} == 413\,100 \simeq (643)^2$ (la variance de l'ensemble est la somme des variances intra et inter).

$\boxed{\mu = 1\,670 \text{ et } \sigma = 643}$

Remarque :

L'écart-type de l'ensemble est 643, l'écart-type intra est 391 et la moyenne des écarts-types est $\bar{\sigma} = \frac{1}{10} 1\,200 + \frac{9}{10} 100 = 210$; on vérifie les inégalités :

$$\boxed{\bar{\sigma} \leq \sigma_{\text{intra}} \leq \sigma.}$$

La deuxième inégalité est une égalité lorsque les moyennes des groupes sont égales (égales alors à la moyenne de l'ensemble).
La première inégalité est une égalité lorsque les écarts-types des groupes sont égaux (égaux alors à l'écart-type de l'ensemble).

Estimation de la moyenne par intervalle de confiance à 95%

Supposons à présent que l'on cherche à estimer le salaire moyen μ à partir d'un échantillon aléatoire de taille 100. On dispose de la liste des 3 000 salariés, avec indication de leur statut (cadre ou employé) et des études antérieures permettent de supposer que l'écart-type des salaires des cadres est égal à 12 fois celui des salaires des opérateurs.

	Effectif	Salaire moyen	Écart-type
Cadres	300		$12 \times \sigma_{\text{ope}}$
Opérateurs	2 700		σ_{ope}
Ensemble	3 000	μ ?	

- un sondage aléatoire simple de taille 100 à probabilités égales avec remise fournit une estimation par intervalle à 95% de confiance de la forme $\hat{\mu} \pm 1.96 \dfrac{\hat{\sigma}}{10}$,
- un sondage aléatoire stratifié proportionnel (10 cadres et 90 opérateurs) fournit une estimation par intervalle à 95% de confiance de la forme $\hat{\mu} \pm 1.96 \dfrac{\hat{\sigma}_{\text{intra}}}{10}$,
- un sondage aléatoire stratifié optimal (57 cadres et 43 opérateurs) fournit une estimation par intervalle à 95% de confiance de la forme $\hat{\mu} \pm 1.96 \dfrac{\hat{\bar{\sigma}}}{10}$,

soit respectivement $\hat{\mu} \pm 126$, $\hat{\mu} \pm 77$ et $\hat{\mu} \pm 41$ en reprenant les valeurs données précédemment au lieu de leur estimation.